ECON 128: Data Science

Claremont McKenna College

Profs. A. Vossmeyer & M. Salloum Due: 9/11/2017

Assignment 1

1 Description

This assignment emphasizes the importance of the first phase of data analytics, which is data understanding and preparation. Before building any models or analyzing your data, its a good idea to inspect the data.

Inspecting your data is a good way to find abnormalities and peculiarities. You might find that your dataset contains many missing values for a particular feature, or that your dataset does not follow a normal distribution.

The purpose of this assignment is to introduce you to useful tools for loading, cleaning and visualizing datasets. The homework assignment is found on the course github page: `https://github.com/msalloum/econ128/tree/master/Homeworks/HW1`.

Before proceeding, lets ensure you have the right tools. We will use the Docker to install the required Data Science libraries. The Docker install instructions are found here: `https://github.com/msalloum/econ128/blob/master/DockerTutorial.pdf`.

For this assignment you will experiment with Python. NumPy, and Pandas in order to perform some basic data preprocessing and analysis tasks.

You will work with a modified subset of a real data set of customer for a bank. The data is provided in a CSV formatted file with the first row containing the attribute names. The description of the the different fields in the data are provided below:

The marketing department of a financial firm keeps records on customers, including demographic information and, number of type of accounts. When launching a new product, such as a "Personal Equity Plan" (PEP), a direct mail piece or a targeted email, advertising the product, is sent to existing customers, and a record kept as to whether that customer responded and bought the product. Based on this database of prior cases, the managers decide to use data mining techniques to build customer profile models in order to predict the behavior of future customers.

The data is contained in the file bank_data.csv. Each record is a customer description where the "pep" field indicates whether or not that customer has purchased a PEP. For classification problems, this field is used as the target attribute (with "YES" and "NO") as class labels.

The data contains the following fields:

1. id a unique identification number (categorical, str)
2. age age of customer in years (numeric, int)
3. income income of customer (numeric, float)
4. children number of children (numeric, int)
5. gender MALE / FEMALE
6. region INNER_CITY/RURAL/SUBURBAN/TOWN
7. married Customer married (YES/NO)
8. car Customer owns one or more cars (YES/NO)
9. save_acct Customer has a savings account (YES/NO)
10. current_acct Customer has a current checking account (YES/NO)
11. mortgage Customer have a mortgage (YES/NO)
12. pep Customer purchased a PEP, Personal Equity Plan (YES/NO)

You must only use Python, NumPy, Pandas, Matplotlib to perform the tasks for this assignment.

1. Explore the general characteristics of the data as a whole: examine the means, standard deviations, and other statistics associated with the numerical attributes; show the distributions of values associated with categorical attributes; etc.
2. Suppose that because of the bank is particularly interested in customers who buy the PEP (Personal Equity Plan) product. Compare and contrast the subsets of customers who buy and don't buy the PEP. Compute summaries (as in part 1) of the selected data with respect to all other attributes. Can you observe any significant differences between these segments of customers? Discuss your observations.

3. Discretize the age attribute into 3 categories (corresponding to "young", "mid-age", and "old"). [Do not change the original age attribute in the table.]

4. Using Matplotlib library and/or ploting capabilties of Pandas, create a scatter plot of the Income attribute relative to Age. Be sure that your plot contains appropriate labels for the axes. Do these variables seem correlated?

5. Create histograms for Income (using 9 bins) and Age (using 15 bins).

2 Submission

Please submit the homework files via Sakai. Alternatively, for 5 extra points, create a github account and a repository called ECON128, and push your homework solution to that ECON128/HW1. You must still submit the repository location via SAKAI is you choose to do the extra credit.